Social Media Text Normalization

Thet Thet Zin
University of Computer Studies (Thaton)
ttzucsy@gmail.com

Abstract

Recent years some researchers interested in text normalization over social media, as the informal writing styles found in Twitter and other social media data. These informal texts often cause problems for Natural Language processing applications such as various mining research or translation on social media data. Today Facebook supports English translation of post and status for Myanmar Language. However, Most of the translation is not relevant for Myanmar words meaning. Complex nature of Myanmar language's syntactic structure, informal writing style, slang words and spelling mistakes are challenge in social media text translation work. This paper proposed text normalization that can be deployed as a preprocessing step for opinion mining, machine translation and various Natural Language Processing (NLP) applications to handle social media text. There are three steps in this work: Firstly, candidate words for normalization are selected from the collected raw dataset. In this case, Out-Of-Vocabulary (OOV) words are extracted for normalization. However, not all OOV words need to be normalized. Therefore, ill-formed words are detected from OOV words list for normalization. Second, slang words dictionary is generated for this work. Third, text similarity methods are applied to ill-formed words for normalization. Evaluation will be done on translation by applying normalization in pre-processing step. For translation, Myanmar-English machine translation [14] is used. The experimental results improve by applying proposed normalization to the translation work especially for social media text.

Keywords- informal text, social media, normalization, Out-Of-Vocabulary word (OOV), translation

1. Introduction

Now, nearly all people use user-oriented media such as social networking sites, blogs and micro blogging services. This led to a rapid increase in the need to understand casual written style, which often does not conform to rules of spelling, grammar and punctuation. Social media text is usually very noisy and contains a lot of typos, ad-hoc abbreviations, phonetic substitutions, customized abbreviations and slang language. The quality of text varies significantly ranging from high quality newswire-like text to meaningless strings. In Myanmar, mix usage of emotional voice and formal words change the meaning of the phrase. This is the big issue for translation especially for word-level translation work (eg. "ကောင်းတယ်ဝိုး"-"satire").This example cannot translate directly. Formal meaning of front word ("ကောင်းတယ်"-"good") is good. However, the meaning of whole word ("ကောင်းတယ်ဝိုး"-"satire") is satire to other. By combining the word ("ကောင်းတယ်"-"good") with the word ("ဝိုး"-informal word), become negative meaning of the word ("ကောင်းတယ်"-"good"). To handle this case, slang word or informal word dictionary is needed. Moreover, some write English pronunciation using Myanmar words: (eg. "တူဒေးမီနူး"- "today menu", wrong translated word is "nephew menu"). Therefore, social media text is often unsuitable as data for NLP tasks such as opinion mining, information retrieval and machine translation due to the irregularity of the language feature. Although average sentence length of social media text is small and generally commented on the posted text, it is not easier to find out the related context. Since social media text includes informal text, slang words, grammar and syntactic errors in the text. Moreover, some informal words have indirect meaning. To handle this case, informal or slang words dictionary and normalization process is needed to capture actual user's opinion for opinion mining.

In previous work [13] some way of preprocessing on the comments data is proposed to produce clean data. Aim in this paper is to normalize some ill-formed words such as multiword expression ("အောင်စေစေစေ"- "be successful") which cannot be solved in previous preprocessing work. Some normalization tasks are similarities with spell checking but differs in that ill-formed in text. Spell checker is also needed to normalize for machine translation. Nevertheless, ill-formed words or slang words like (eg. "ကွီး၊ မတ်မတ် is slang word. Formal words is "ကိုကြီး၊ မမ- brother,sister") tend to be considered beyond the scope of spell checking. In addition, the detection of informal words is difficult due to noisy context. The objective of this work is to detect ill-formed word and normalize to standard Myanmar word for translation. Similarity method is applied to OOV words. Category of OOV will discuss in the next section. In this approach a list

of candidate word for normalization is generated firstly. Then slang word dictionary for Myanmar language is generated for normalization. Finally, similarity calculation is done between ill-formed words and candidate words. Proposed method supports to improve F-score and BLEU score in translation work.

Contributions in this paper are as follows: (1) studying the OOV word distribution of text and analyze different sources of non-standard orthography in data; (2) generating a slang words dictionary based on social media text; (3) detecting ill-formed words for normalization work exploits dictionary lookup and word similarity without requiring annotated data; (4) demonstrating the method better support for translation over social media text.

2. Related Work

Research aimed at the specific problem of normalizing casual Myanmar language is relatively rare. Some researcher fixed this problem by using NLP tools to social data. NLP tools for Myanmar language are rare in present time. The normalization approach is especially attractive as a pre-processing step for applications, which rely on key word match or word frequency statistics. For example, "အောင်စေ အောင်စေစေစေစေ အောင်စေစစစစစ- "be successful") all attested in a Facebook comments corpus – have the standard form "အောင်စေ"- "be successful"; by normalizing these types to their standard form, better coverage can be achieved for keyword-based methods, and better word frequency estimates can be obtained.

The range of problems presented by user-generated content in online sources go beyond simple spelling correction; other problems include rapidly changing out-of-vocabulary slang, short-forms and acronyms, punctuation errors or omissions, phonetic spelling, misspelling for verbal effect and other intentional misspelling and recognition of out-of vocabulary named entities [2]. To discover the sequential dialogue structure of open-topic conversation in Twitter, [3] proposed unsupervised based conversation model. They compared Bayesian inference to Expectation-Maximization (EM) on conversation ordering task, showing a clear advantage of Bayesian methods. Hany and Arul [4] propose another unsupervised learning of the normalization equivalences from unlabeled text. They presented contextual graph random walks for social text normalization. Their proposed system based on constructing a lattice from possible normalization candidates and finding the best normalization sequence according to an n-gram language model using a Viterbi decoder. In addition, used random walks on a contextual similarity graph constructed form n-gram sequences on large unlabeled text corpus. They evaluated the approach on the normalization task as well as machine translation task. They figured out some limitations in normalization task and did not consider for mixed usage of words (eg, text include Myanmar and English words). Qi and other researchers [5] proposed Chinese-English mixed text normalization work. Experimental results on a manually annotated micro blog dataset demonstrate the effectiveness of their proposed method. From the results, this method can significantly benefit other NLP tasks in processing mixed usage of Chinese and English. Some researchers divide the text normalization problem into two sub-categories: word-based and character-based normalization. The word-based normalization turns non-standard words such as slang, acronyms and phonetic substantiation into standard dictionary words. Character-based normalization transforms the raw text through substituting the irregularly used characters with proper ones. Unsystematic usage of Latin alphabets (UULA) is presented by Osman and Ruket on noisy Uyghur text [6]. UULA normalization is character-level normalization. The noisy channel model and the neural encoder-decoder model are proposed and compared as normalizing methods. The noisy channel model views the problem as a spell-checking problem, while the neural encoder-decoder model views it as a machine translation problem. Both of them return highly accurate results on restoration and recommendation tasks on the synthetic dataset. However, their accuracy on real dataset would benefit from further improvement. To improve their performance on the real dataset, one possible strategy is to consider other noisy factors appearing in the real dataset.

Now especially at the social media in Myanmar, most users use informal writing style and appearing many slang words. Grammar and syntactic mistake also found in social media text. These cause issue for translation processes. Therefore, normalization for Myanmar social media text is needed. According to my knowledge, it is very rarely related research work in this area.

3. Data Analysis

As already described above, most of users, write status and comments using informal text in social media. This data need to be normalized for further processing. In this section, the dataset is examined for better understanding of the nature of data collected from Facebook. According to analysis, they use abbreviation (short form or acronym), slang word, mix typing usage (Myanmar and English word eg. (tmrသွားမှာလား- will tomorrow go?), multiword expressions, emotion icons and syntactic mistake. During the present time, many slang words in Myanmar language appeared via Facebook. Data from Facebook is collected by using Facebook API. Firstly, data is analysis into two parts formal and informal text. In the informal text category, abbreviation (short form and acronym), non-dictionary slang words, multiword expressions, mixed usage of two languages, orthographic mistakes, omission of vocabulary, combining two or three

words to one slang word and further categories: Named entity, swear-word censor avoidance, emotion icon have been included. For this work, Facebook status data is extracted from 1st June 2018 to 1st July 2018. There are 20,897 sentences with length between 20 and 35. To analysis formal and informal text percentage in collected data, we selected 1,000 sentences from dataset randomly. 68% of selected sentence use formal writing style and 32% are using informal style. Most of the informal texts are phonetic substantiation into standard dictionary words. The detail analyze of informal text is described in the table.

Table1. **Category of informal text**

Category	Percent	Example
Abbreviation (short form or acronym)	5%	ဝကခ (ဝန်ကြီးချုပ်) ၊ မလမ (မော်လမြိုင်)
Omission of vocabulary	20%	ခေး(ကလေး) ၊ ကွီး(ကိုကြီး) ၊ မွီးဂါး (မုန့်ဟင်းခါး)
Mix typing usage	5%	Trmလာမယ် ၊ okလေ
Multiword expression	10%	အောင်မြင်ပါစေစစစစ ၊ အောင်မြင်ပါစေ!!!!
Emotion icons	10%	☺ :P
Orthographic mistakes	20%	**ဆိုက်**ထားတဲ့အပင်လေး
Myanglish	10%	kaung par pi
Slang word	10%	အယ်လယ် ၊ လန်းချက်
Others (** named entity and Swearword Censor **)	10%	အောင်မင်္ဂလာအဝေးပြေး၊ စောက်သုံးမကျလိုက်တာ၊ ငိုငိ ၊ ခီခီ

Spelling checker handles orthographic mistakes in the text. Now, it is not possible to integrate Myanmar word spelling checker in the process. Myanglish (using English words for Myanmar words pronunciation: eg 'နေကောင်းလား-how are you?'- *nay kg lar?*) words are difficult for normalization because writing style of one different from another. Moreover, detecting and analyzing emotion icons will do separate research in the future. Other category includes named entity and Swearword Censor Avoidance. Therefore, these four categories are out of this paper.

4. Ill-formed Words Detection and Normalization

Detecting ill-formed words for normalization is a challenging problem especially in social text for many reasons. First, it is not straightforward to define the Out-of-Vocabulary (OOV) words. Traditionally, an OOV word is defined as a word that does not exist in the vocabulary of a given system. However, this definition is not adequate for the social media text, which has a very dynamic nature. Many words and named entities that do not exist in a given vocabulary should not be considered for normalization. Moreover, same OOV word may have much appropriate normalization depending on the context and on the domain. Therefore, analysis for words for normalization is difficult in social media text. In this paper, four steps are proposed for detecting candidate words for normalization.

First, blank space and punctuation are removing from n-gram words sequence. Myanmar sentences are segmented by using Myanmar syllabus segmenter developed by knowledge engineering major students of University of Information Technology (UIT). Present time, this segmentation tool cannot upload into the university website. Accuracy of this tool is reported in previous work [13] and the project book of this tool can get in UIT's FCS department. The longest matching n-gram is applied to segmented Myanmar sentence. Tri-gram is the best for this work.

Second, Myanmar-English bilingual corpus for machine translation is used for this work. Formal Myanmar-English corpus are created since previous work [14] on open domain. There are 61,824 Myanmar words and 56,263 English words. Every n-gram word is searched in the corpus. Some words do not have individual meaning but they have meaning by combining other surrounding words. Therefore, longest matching is used in this work.

Third, Myanmar words have many prefix, suffix, counting words and stop words. There are also remove from OOV words list. There are 603 prefix, suffix and counting words [13]. After doing above three steps, the remaining words may be OOV words or candidate words for normalization.

Fourth, two similarity methods and slang words dictionary are used to calculate similarity value in candidate words and ill-formed words. Firstly, slang words are extracted by using created slang words dictionary. After extracting slang words, normalization is applied to these words. Format of slang words dictionary is shown in table2. For example; the word 'မွီးဂါး' is normalized to 'မုန့်ဟင်းခါး (Myanmar traditional food)' using slang word dictionary. After searching slang words, two ways of similarity are calculated on remaining OOV words in the sentence. Words can be similar in two ways lexically or semantically.

Words are similar lexically if they have a similar character sequence. String-based n-gram similarity is used for lexical similarity based on the number of shared n-gram. *X* is the candidate words and *Y* is the canonical form in the dictionary. Similarity measure is calculated using the number of shared n-gram between two words. If value is greater than or equal to 0.6, it assumes the two words are similar lexically and normalized the candidate word with the canonical form. For example the candidate word ('အောင်စေစေစေစေ-be successful') is normalized to the

canonical form (‘အောင်စေ’). If lexical similarity value of the word is less than 0.6, semantic similarity for this word is calculated.

Semantic similarity is calculated for the words which are not similar lexically in the sentence. Words are similar semantically if they have the same words, used in the same way, used in the same context and one is a type of another. In this case, surrounding words of the OOV should be considered. *X* is the candidate words and *Y* is the canonical form of the word. Probability $i_n(X, Y)$ is calculated by using surrounding words of the candidate words in the sentence. For example; in the sentence “လေတဟုန်းဟုန်းတိုက်တယ် - The wind roared.” the word “တဟုန်းဟုန်း-roared” is OOV words and lexically similar word are not contains in the dictionary. But the dictionary contains the word “ဝုန်းခနဲ”. Probability of $i_n(X, Y)$ is calculated by adding surrounding n-grams words combination probability for two sentences “*လေတဟုန်းဟုန်းတိုက်တယ်*” and “*လေဝုန်းခနဲတိုက်တယ်*” – “The wind roared” and the corpus. If the probability value is greater than 0.5, it assumes that the two sentences has nearly the same content. One of the issues of this calculation is that the probability value totally depends on corpus size and sentences in it. This word does not need to normalize for translation. But it can reduce OOV words for translation process. Another challenge in this similarity calculation is the words order in the sentence. Eg. “*တဟုန်းဟုန်း**လေ**တိုက် တယ်*” and “***လေ**ဝုန်းခနဲတိုက်တယ်*”. At the present time, one direction combination probability is done in this work. To analysis detection of ill-formed words from OOV, we selected 1,000 sentences randomly. Analysis result shows in figure 1.

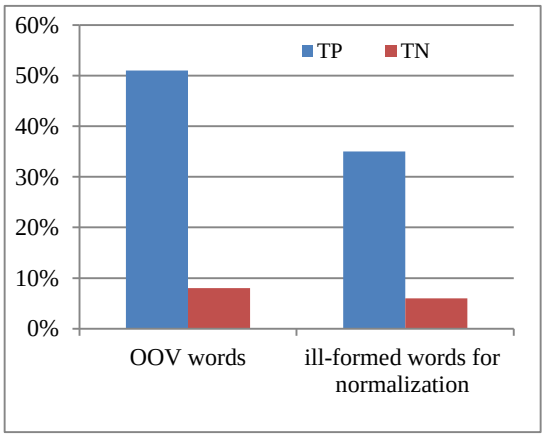

Figure1. Analysis of ill-formed words detection

According to analysis in figure 1, true positive is one that detects the condition (OOV words and ill-formed words) when the condition is present. True negative result is one that does not detect the condition when the condition is absent. 8% of true negative in OOV words include named entity and spelling error. Sometime, Myanmar words are similar semantically or word ordering (eg. အသုပ်စုံ၊အစုံသုပ်၊အသုပ် has the same translation word ‘salad’) but the system detects these words as OOV for domain. To analysis semantic similarity words, sometime it should consider the surrounding words also. 6% of true negative in ill-formed words include substitution phonic and slang words which do not include in dictionary (eg. ‘အကွက်တွေ’ is slang word. Formal word is ‘လှည့်ကွက်’. Translated word is ‘trick’). This is difficult to handle all slang words, which is used in social media. In the future, the dictionary will be perfect more than present time.

5. Normalization Lexicon Generation

We collected social media data from Facebook to generate normalization lexicon. Facebook statuses are collected for one month using Facebook API. There are 10,234 Myanmar sentences. Average words length of these sentences is 28. This paper uses a manually compiled and verified database, currently of a total of 805 entries. This amount is very small for normalization. These entries are either single words or phrases. At present time length of phrase entries are sets of two or three words. Each entry has been taken from separate sentences training data collected from Facebook status and comments. Database entries comprise of three columns: “the casual Myanmar word”, “regular word” (the corresponding dictionary Myanmar word) and “category”. One standard word has many relevant slang words. Database construction is an ongoing project, and intends to improve its coverage and quality further. Later, we will use unsupervised approach for generation for lexicon instead of manually compiled. Format of slang word dictionary is shown in table2.

Table 2 .Format of Myanmar slang dictionary

Casual word	Regular word	Category
ဝကခ၊ မလမ	ဝန်ကြီးချုပ်(prime minister)၊ မော်လမြိုင်(mawlamyine)	Abbreviation
မိုးဂါး	မုန့်ဟင်းခါး (Myanmar traditional food)	Omission of vocabulary
Today မီးနူး	တူဒေးမီးနူး(Today menu)	Mix typing usage
လန်းချက်	လှသည် (beautiful) မိုက်သည် (cool)	slang word

6. Experiential Result

We constructed a test set of 1,000 sentences with average sentences length 10 words are collected from social media, which are separated from training data set.

Furthermore, a test set is developed for evaluating the effect of the normalization process when used as a preprocessing step for translation work. A test set for human evaluation and BLEU scores. Human evaluation results are shown in figure 2. For translation, we used Myanmar-English translation proposed in [14]. We prepared translation reference for test set under the guideline of English lecturer. Precision and recall of words' translation are calculated by the following way.

$$precision \quad \frac{correct\_words}{output\_lenght}$$

$$recall \quad \frac{correct\_words}{reference\_lenght}$$

$$F-measure \quad \frac{precision * recall}{(precision + recall)/2}$$

Table3. Example of translated sentence

Myanmar sentence	ဂရုတစိုက်နားထောင်ပါမတ်မတ်
Reference:	listen carefully sister
Translation without normalization	listen carefully *march march*
Translation with normalization	listen carefully sister

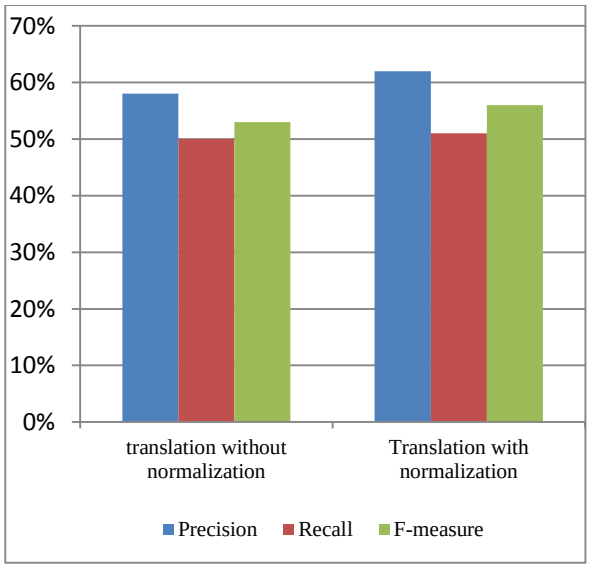

Figure2. Evaluation results for normalization

Most of the sentences can be translated in many acceptable forms. Thus, more than one reference sentences should be considered. One reference is considering for the results and ignores word order in the translated sentence. BLEU is a score for comparing a candidate translation of text to one or more reference translations. Higher numbers correspond to better translations. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. Some translated output is much too short, thus boosting precision, and BLEU doesn't have recall. We evaluate the translation based on 4-grams BLEU scores evaluation. The results are shown in the following table.

Table4. 4-gram BLEU score

	BLEU
without normalization	0.296
with normalization	0.373

Analysis on 1,000 sentences, 52% of precision for these sentences comes in human evaluation. This meaning that false positive rate increases in testing dataset. Ill-formed words detection has some errors. Recall on this case is higher than precision. This means that training data for this test set is reasonable enough. In the translation work, precision is higher than recall. Increasing the amount of training data will affect to the performance positively especially the recall. We also test translation without applying normalization process. The results show that translation uses normalization as a preprocessing step for a machine translation, which improved the translation quality by 3% in F-score.

In BLEU score of the translation with normalization, about 75.4% of the overall precision score comes from the uni-grams. 17.5% comes from the bi-grams; 4-grams contribute only 1.3%. The number of longer n-gram matches is smaller compared with shorter n-gram matches. we assume that the human evaluation scores are the most valid then the automatic metrics for only these 1,000 test set.

7. Discussion and Future Work

We first manually analyses the errors in normalization over the test set. There are two categories in error analysis. First is an error in ill-formed words detection. The most frequent in this category is caused by morphological variations, including (1) negations: In Myanmar language, negation is difficult to recognize for further processing because Myanmar has many negation forms. (eg. မရဘူးဟာ၊ ရမယ်လို့မထင်တာ၊ ရကိုမရတာ၊ မသိဘူးမရ ဘူး- all are negation form). It fails to normalize 32% of the negation words in the test set (2) syntactically or semantically ambiguity between words: about 23.5% of the words have ambiguous meaning (3) spelling errors: about: over 30% of the words have spelling errors (4) over 10%

of the words are missing in created slang words dictionary. The second category is false positive rate in OOV words. Some words have same meaning with different forms. This cause occurs in OOV words or ill-formed for normalization. This reduces precision of the translation. We already mention above that the translation uses normalization as a preprocessing step for a machine translation which improved the translation quality by 3% in F-score. Most errors in this case are that in social media text, some words cannot be translated directly. It cannot be translate by only considering surrounding words and sentence structure. For example: အထောင်းမှန်သမျှ-'all pounded food: such as pounded papaya' this word cannot translate English word "right" even through ("မှန်" is "right"). Some errors found in changing slang word to standard word for translation. Because some slang word has different meaning depend on content of the text. For example: (အကွက်တွေမိုက်တယ်- 'nice trick' ၊ သန့်နေတာဘဲ- 'neat and tidy'). Normalization process does not know these words need to normalize for translation work. Some output show that normalization process has done on words but normalized standard word is wrong for translation. To overcome this problem, consideration on content and sentence structure of the text include both normalization and translation processes. Moreover, slang word dictionary will need to powerful than before.

8. Conclusion

In this paper, normalization on a social media text is proposed that can be deployed as a preprocessor for MT to handle social media text. We analyzed the collected data and identified ill-formed words for normalization. Most informal text in social media based on spelling mistake, slang words and substation of phonic. Proposed informal text detection method shows accepted results. However, other experiment and methodology are needed to improve ill-formed word detection. Moreover, slang words database generation is an ongoing project. For effective normalization on social media text, powerful annotation corpus or effective unsupervised method is needed. Some limitations in proposed approach are found by analyzing output results: example mix type usage cause the problem for normalization. As an extension to this work, we will extend the approach to handle named entity and spelling mistake by integration Myanmar named entity recognition and spelling checker to the normalization on social text. Furthermore, the approach can be extended to handle semantically similar words problems for normalization. We hope the best results will outcome in the future.

9. References

[1] E.Clark and K.Araki, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English", *Pacific Association for Computational Linguistics (PACLING 2011)*, Published by Elsevier Ltd. 2011, pp. 1-10.

[2] E.Clark, T.Roberts and K.Araki, "Towards a pre-processing system for casual English annotated with Linguistic and Cultural Information", Proceeding of the fifth IASTED International Conference Computational Intelligence (CI 2010), August 23-25, 2010 Maui, Hawaii,USA.

[3] A. Ritter, C. Cherry and B. Dolan, "Unsupervised Modeling of Twitter Conversations", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, June 2010, Los Angeles, California, pp 172-180.

[4]H. Hassan and A.Menezes, "Social Text Normalization using Contextual Graph Random Walks", Proceedings of the 51st annual meeting of the association for computational linguistics, August 4-9 2013, Sofia, Blugaria, pp 1577-1586.

[5] Q.Zhang, H.Chen and X.Huag, "Chinese-English Mixed Text Normalization", WSDM , February 24-28 2014, New York, USA, Copyright 2014 ACM 978-1-4503-2351-2/14/02, pp 433-42.

[6] O. Tursun and R. Cakici, "Noisy Uyghur Text Normalization", Proceedings of the 3rd Workshop on Noisy User-generated Text, September 7, 2017, Copenhagen, Denmark, pp- 85-93.

[7] T. Baldwin and Y.Li, "An In-depth Analysis of the Effect of Text Normalization in Social Media", Human Language Technologies: The 2015 annual conference of the North American chapter of the ACL, May 31- June 5, 2015, Denver, Colorado, pp 420-429.

[8]B. Han and T. Baldwin, "Lexical Normalization of Short Text Messages: Makn Sens a #twitter ", proceedings of the 49th annual meeting of the association for computational linguistics, June 19-24, 2011, Portland, Oregon, pp 365-378.

[9] C. Henriquez Q and Adolfo Hernandez H, "A Ngram-based Statistical Machine Translation Approach for Text Normalization on Chat-speak Style Communications", CAW 2.0, April 21 2009, Madri, Spain.

[10] S. Rohatgi and M. Zare, "DeepNorm- A Deep learning approach to Text Normalization", ACM ISBN 123-4567-

24-567/08/06, IST597-003 Fall' 17, December 2017, State College, PA, USA.

[11] B. Han, P.Cook and T. Baldwin, "Automatically Constructing a Normalization Dictionary for Microblogs", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, July 12-14 2012, Jeju Island, Korea, pp 421-432.

[12] S. Goyal and Er. Bedi, "SMS Text Normalization Using Hybrid Approach", International Journal of Computer Trends and Technology (IJCTT), Volume 21 Number 2, ISSN: 2231-2803, march 2015, pp126-129.

[13] Zin et al.,"Domain-Specific Sentiment Lexicon for Classification", the First International Conference on Advanced Information Technologies (ICAIT), November 1-2, 2017, Yangon, Myanmar.

[14] T.T.Zin, K.M.Soe and N.L. Thein, "Translation Model of Myanamr Phrases for Statistical Machine Translation", the 2011 seventh international conference on intelligent computing, august 11-14 2011, Zhengzhou, Henan, China, copyright Springer-Verlag Berlin Heidelberg 2011, pp. 235-242.

[15]http://www.ucsy.edu.mm/GoNLP.do